Advances in Artificial Intelligence and Machine Learning; Research 6 (4) 6033-6049 Received 03-04-2026; Accepted 11-08-2026; Published online 18-08-2026

A Text-Guided Cross-Modal Diffusion Framework With Attention-Based Hand-Drawn Sketches for Face Synthesis

Inas Ismael Imran phd202410014@uoitc.edu.iq
Informatics Institute for Postgraduate Studies,
University of Information Technology and Communication (UoITC),
Baghdad, Iraq

Ahmed A. Hashim dr.ahmed.hashim@uoitc.edu.iq
Department of Business Information Technology,
business Informatics college, University of Information Technology and Communication (UoITC),
Baghdad, 10053, Iraq

Corresponding Author: Ahmed A. Hashim

Abstract

Face sketch-to-photo synthesis plays a key role in computer vision with applications in law enforcement, digital entertainment, and human–computer interaction. Existing generative adversarial network-based methods typically face mode collapse, training instability, and poor performance across different sketching styles. This study introduces Sketch-to-Face, a cross-modal diffusion-based model that leverages Stable Diffusion to generate photorealistic faces using sketch-based inpainting. The proposed approach comprises three components: a Sketch Encoder with multiresolution attention that produces CLIP-compatible embeddings from grayscale sketches, Cross-Modal Fusion module employing bidirectional attention to bridge sketch spatial features with text semantic features, and Automatic Mask Generator with learnable refinement for adaptive inpainting guidance. A low-rank adaptation was applied to fine-tune the U-Net, reducing trainable parameters to 20.5 M while retaining the 860M-parameter backbone. Experiments on the Person Face Sketches dataset (21K+ pairs) show stable convergence and are evaluated using SSIM, LPIPS, FID, identity similarity, runtime, and GPU memory. The results demonstrate a favorable quality–efficiency trade-off compared with text-and-sketch diffusion baselines, while qualitative results indicate plausible facial structure and visual realism.

Keywords: Face synthesis, Sketch-to-photo, Diffusion models, Cross-modal fusion, LoRA, Attention mechanism, Inpainting.

1. INTRODUCTION

Face sketch-to-photo synthesis is a long-standing problem in computer vision that aims to generate highly realistic facial images from hand-drawn or forensic sketches [1]. The practical uses of this

Citation: Inas Ismael Imran and Ahmed A. Hashim. A Text-Guided Cross-Modal Diffusion Framework With Attention-Based Hand-Drawn Sketches for Face Synthesis. **Advances in Artificial Intelligence and Machine Learning**. 2026;6(4):334. https://dx.doi.org/10.54364/AAIML.2026.64334

task are immense, especially in law enforcement (to identify suspects) [2], in the field of digital entertainment (character creation) and human-computer interaction [3]. The main problem is how to make the large domain gap between the abstract sketch representations and the rich texture, color, and lighting information found in natural photographs. The initial models of face sketch synthesis rely on exemplar-based models that obtained and concatenated patches on a source set [1, 3]. Even though these methods could generate reasonable results, they were restricted by the size and diversity of the reference database and often formed artifacts at patch boundaries. Image synthesis has also changed the approach introducing deep learning and, more precisely, generative adversarial networks (GANs) [4]. The first to demonstrate that conditional GANs can be trained to learn image-to-image translation mappings was Pix2Pix [5]. The field has since been extended by other researchers such as the developers of Cycle GAN [6] and Pix2PixHD [7]. Nevertheless, the mode collapse, issues with trainability, and the notorious issue of generators and discriminators optimization tradeoffs still afflict GAN-based algorithms [8, 9]. Recently, diffusion models [10, 11], the powerful alternatives to GANs, have become successfully able to deliver the state-of-the-art results in the quality of the generated images [12] due to being of superior quality compared to GANs.

Latent diffusion models (LDMs) are highly successful in high-resolution synthesis with fair computational expenses. They operate by executing the diffusion procedure in a latent space that has been compressed instead of the original high-resolution image space. Stable Diffusion, which follows the LDM architecture, has demonstrated good results in other tasks such as text-to-image generation [13, 14], and image inpainting [15]. Integrating textual conditioning with CLIP embeddings has been applied to these models to increase controllability [16]. In spite of these developments, the field of sketch-to-face synthesis using diffusion models has unique peculiarities. In contrast, sketches, as compared to natural images, tend to be in black and white, lack any texture data, and are highly varied in the visual styles and details of outlines. The existing conditional diffusion models such as ControlNet [17] and sketch-guided ones [18] are partial solutions, but they can be costly to compute and cannot encode appropriate facial details stored in sketches. To resolve these limitations, this paper presents a new framework, Sketch to Face, with three main innovations:

1. A sketch encoder with multiresolution attention mechanisms that transform grayscale sketch inputs into CLIP-compatible embedding sequences, enabling seamless integration with the pretrained text-conditioned diffusion pipeline (Section 3.2).
2. A cross-modal fusion module that incorporates the use of bidirectional attention to combine sketch spatial features with text semantic features enables the model to utilize both the visual structure and textual description (Section 3.3).
3. A memory-efficient training strategy incorporating low-rank adaptation (LoRA) [19], automatic mixed precision (AMP) [20], and gradient check pointing [21], which reduces trainable parameters to 20.5 M and limits GPU memory usage to 3.92 GB (Section 4.5).

2. RELATED WORK

Synthesis of face sketches has been widely researched over the last 20 years. The first contribution to the field was made by Wang and Tang (2009) [1], who presented eigen-transform methodology,

which trained a linear transform between the photo space and sketch space using the CUHK Face Sketch database. This was generalized by Zhang et al. (2011) [2], with coupled information-theoretic encoding, and Gao et al. (2012) [3], with sparse representation-based sketch-photo retrieval. Zhang et al. (2018) [22], introduced a coarse-to-fine method, which gradually improved the generated photos. A general deep-learning benchmark for sketch synthesis evaluation was proposed by Peng et al. (2018) [23]. Although these methods were basic, they were based on handcrafted properties [24], or shallow networks, which restricted their ability to learn complicated sketch-to-photo mappings. GANs transformed the process of image generation by positioning generation as an adversarial game between a generator and a discriminator [4]. Translation conditional GANs presented in Pix2Pix [5] showed that paired training examples can be used to learn image-to-image translations, such as sketch-to-photo translations. Cycle GAN [6] eased the paired data condition using cycle-consistency losses. Multiscale Pix2PixHD [7] trained conditional generators and discriminators at a high resolution. StyleGAN [8] and StyleGAN2 [9] introduced style-based architectures that enabled unprecedented control over generated face attributes. pSp [25] further explored encoding images into the StyleGAN latent space for image-to-image translation. While GAN-based approaches achieve impressive results, they remain sensitive to hyperparameter tuning and are prone to mode collapse [4]. Denoising diffusion probabilistic models (DDPMs) [10] generate images by learning to reverse a gradual noising process. Song et al. (2020) [11], accelerated sampling through implicit diffusion. Dhariwal and Nichol (2021) [12], demonstrated that diffusion models can outperform GANs in terms of image quality. Rombach et al. (2022) [26], proposed LDMs that operate in a compressed latent space, which dramatically reduces computational costs. Saharia et al. (2022) [13], incorporated Imagen for photorealistic text-to-image synthesis, while DALL-E 2 [14] combined CLIP with diffusion for hierarchical generation. For conditional generation, classifier-free guidance [27] enables a trade-off between diversity and fidelity without a separate classifier. ControlNet [17] adds spatial conditioning to pretrained diffusion models, and sketch-guided diffusion [18] demonstrates sketch-conditioned synthesis. Our work builds on the Stable Diffusion inpainting model and introduces a novel sketch-specific conditioning (see TABLE 1).

Table 1: Comprehensive comparison of related work in sketch-to-face synthesis and conditional image generation.

Method	Year	Architecture	Paired Data	Resolution	Parms (M)	Cross-Modal	Memory Eff.
Wang & Tang [1]	2009	Eigen-transform	Yes	200 × 250	< 1	No	Yes
Zhang et al. [2]	2011	Info-theoretic	Yes	200×250	< 1	No	Yes
Gao et al. [3]	2012	Sparse repr.	Yes	200× 250	< 1	No	Yes
Pix2Pix [5]	2017	Cond. GAN	Yes	256 ×256	~ 54	No	No
CycleGAN [6]	2017	Cycle GAN	No	256×256	~ 28	No	No
Pix2PixHD [7]	2018	Multi-scale GAN	Yes	1024× 1024	~ 183	No	No
CartoonGAN[28]	2018	GAN	No	256 ×256	~ 11	No	No
Zhang et al. [22]	2018	Coarse-to-fine GAN	Yes	256× 256	~ 40	No	No
Peng et al. [23]	2018	Deep CNN	Yes	256 ×256	~ 25	No	No
pSp [25]	2021	StyleGAN encoder	Yes	1024×1024	~ 70	No	No
ControlNet [17]	2023	Cond. diffu-sion	Yes	512 × 512	~ 361	Partial	No
Sketch-guided [18]	2023	Guided diffusion	No	512 ×512	~ 860	Partial	No
Sketch-to-Face (Ours)	2026	Cross-modal diff.	Yes	256×256	20.5	Yes	Yes

3. METHODOLOGY

FIGURE 1 depicts a text-guided sketch-to-face architecture trained on the Person Face Sketches dataset, which consists of 21,334 sketch- photo pairs in train, validation and test splits. The 256×256 sketch is embedded into features space, while the text prompt is encoded via CLIP into vector representation such as [B, 77, 768], where B is batch size, 77 is sequence length, and 768 is feature dimension. the modalities are integrated, then a spatial mask m [1,32,32] facilities region enhancement before the U-Net + LoRA conduct iterative denoising via diffusion across 50 steps with guidance confienciat s = 7.5. moreover, the VAE decoder reconstructs the latent into a 3×256×256 RGB facial output, while values such as 1.78M, 123M, 17.1M, and 83.7M represent the parameter sizes of the core modules in the architecture. The Conditioning Pathway encodes the input sketch and text prompts into CLIP-compatible embeddings, which are merged via bidirectional cross-attention in the Cross-Modal Fusion module. The fused embedding e f conditions the LoRA-adapted U-Net during Iterative Denoising, whereas the mask generator provides Spatial Guidance derived from the sketch structure. Solid-bordered modules are trainable (20.5 M parameters) and dashed-bordered modules are frozen pretrained components. The VAE decoder maps the denoised latent z0 to the final photorealistic face x^.

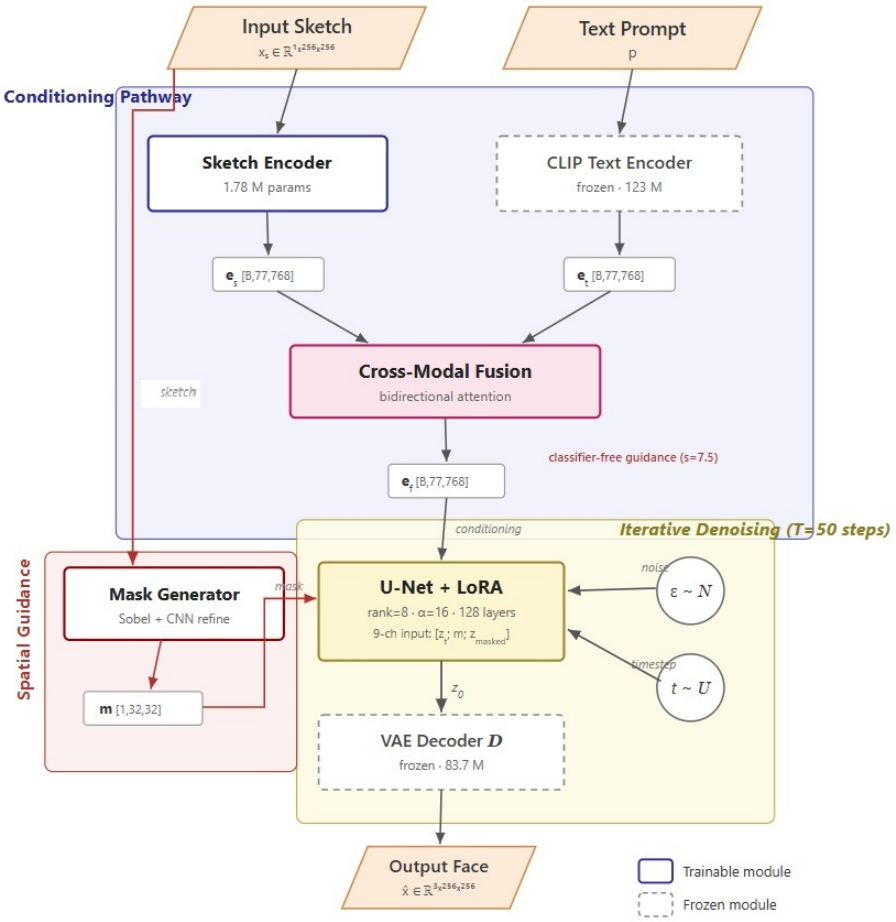

Figure 1: Overall architecture of the proposed Sketch-to-Face framework (vertical flow)

3.1 Problem Formulation

Given a grayscale face sketch xs$\in R^1 \times H \times W$ and an optional text prompt p, our goal is to synthesis a photorealistic face image x^ $\in R^{3\times H\times W}$ that preserves the facial structure and identity encoded by the sketch. We formulate this as a conditional generation problem.

$x^ = G(x_s, p; \theta)$ (1)

where G denotes our Sketch-to-Face pipeline parameterized by θ, which consists of four trainable components: the sketch encoder Es, the cross-modal fusion module F, the mask generator M, and the LoRA-adapted U-Net $\epsilon\theta$ "(see FIGURE 1)".

3.2 Sketch Encoder

The sketch encoder Es transforms a single-channel sketch image into a sequence of embeddings compatible with the CLIP text encoder output space (RB×77×768). This design enables direct substitution of text conditioning without modifying the pretrained U-Net architecture.

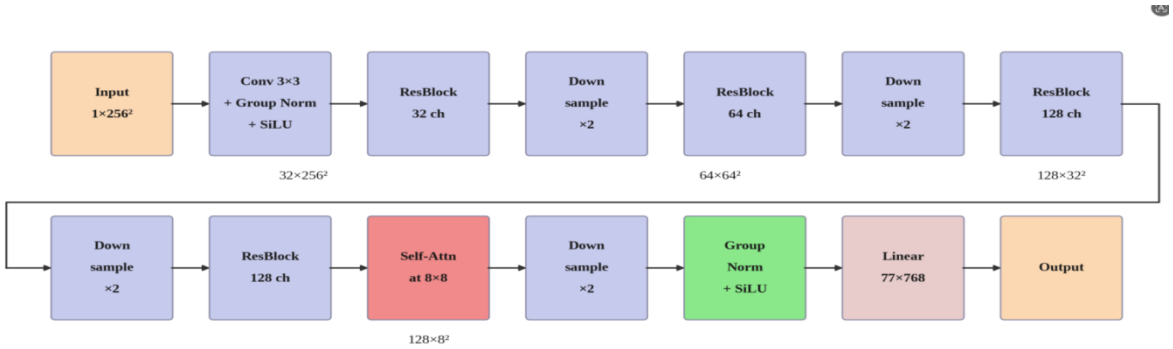

Figure 2: Detailed architecture of the sketch encoder. The encoder progressively downsamples

The input sketch through residual blocks with increasing channel dimensions, applies self-attention at the lowest resolution (8 × 8), and projects the features to a sequence of 77 tokens with 768 dimensions, matching the CLIP text encoder output format (see FIGURE 2).

3.3 Cross-Modal Fusion

The cross-modal fusion module F combines sketch embedding, es, with text embedding, et, (obtained from the frozen CLIP text encoder) through bidirectional cross-attention. This is a key innovation in our approach, as it allows the model to leverage both geometric information from the sketch and semantic guidance from text prompts. The fusion module consisted of L = 1 bidirectional fusion block. Each fusion block performs the following operations:

e’s = es + CrossAttn(LN(es), LN(et)), (2)

e’t = et + CrossAttn(LN(et), LN(es)), (3)

e’s’ = e’s + FFN(LN(e’s)),e’t’ = e’t + FFN(LN(e’t)), (4)

where LN denotes layer normalization, CrossAttn is multihead cross-attention with 8 heads, andFFN is a feed-forward network with GELU activation and an expansion factor of four.

3.3.1 Distinction From Control Net Framework

This paper used a trainable bidirectional sketch–text fusion mechanism for text-guided sketch-to-face synthesis. Unlike ControlNet-style conditioning, where the external sketch mainly acts as

a spatial control signal for a pretrained diffusion model, the proposed method learns an explicit interaction between the sketch representation and the text representation before the denoising stage. In this design, the sketch embedding is semantically enriched by the text prompt, while the text embedding is simultaneously aligned with the geometric structure of the sketch. This produces a unified semantic–structural conditioning representation that is then passed to a LoRA-adapted U-Net. Therefore, the novelty lies in integrating bidirectional cross-modal alignment, sketch-guided facial structure preservation, and parameter-efficient LoRA adaptation within a single text-guided sketch-to-face generation framework.

3.4 Low-Rank Adaptation (LoRA)

To enable memory-efficient fine-tuning, we applied LoRA [19] to all attention projection layers in the pretrained U-Net. For a pre-trained weight matrix W0 ∈Rd×k, LoRAreparameterizes the weight update as:

$$\mathbf{W}=\mathbf{W}_0+\frac{\alpha}{r}\mathbf{BA}, \quad (5)$$

where $\mathbf{A}\in\mathrm{R}^{r\times k}$ and $\mathbf{B}\in\mathrm{R}^{d\times r}$ are the low-rank matrices with rank r = 8, and $\alpha = 16$ is the scaling factor. A is initialized with K aiming uniform , and B is initialized to zero such that the initial output is equal to that of the pretrained model. We applied LoRA to eight attention layers in U-Net, with all the LoRA parameters maintained in FP32 for gradient stability during mixed-precision training (see FIGURE 3).

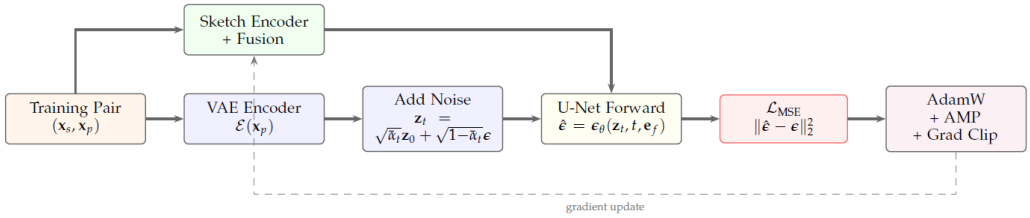

Figure 3: Training pipeline of Sketch to Face.

During training, the target photo is encoded to latents, noise is added at a random timestep t ∈ [0, 800), and the U-Net predicts the noise conditioned on fused sketch-text embeddings. The MSE loss between predicted and actual noise drives gradient updates through all trainable components

3.5 Training Procedure

Algorithm 1 presents the complete training procedure. Sketch-to-Face Training Procedure

Require: Training set D = $\{(X_s^{(i)}, X_p^{(i)}, p(i))\}^N$, $_{i=1}$ pre-trained SD-inpainting model, learning rate η, epochs E, gradient accumulation steps G

Ensure: Trained pipeline parameters θ

1: Initialize Sketch Encoder E_s,Fusion F, Mask Generator M

2: Apply LoRA (r = 8, α = 16) to U-Net attention layers

3: $\theta \leftarrow$ { Es, F,M, LoRA params}

4: optimizer AdamW$\leftarrow(\theta, \eta$, weight _ decay= 0.01)

5: scheduler $\leftarrow$ LinearWarmup (200 steps) + Constant

6: for epoch = 1 to E do

7: for batch (x_s, x_p, p) in DataLoader(D) do

8: $e_s \leftarrow$ Es(x_s) ? Encode sketch

9: $e_t \leftarrow$ CLIP(p) ? Encode text (frozen)

10: $e_f \leftarrow F(e_s, e_t)$? Cross-modal fusion

11: $z_0 \leftarrow E(x_p) \cdot S_{vae}$? Encode photo to latent

12: m $\leftarrow$ M(x_s) $\leftarrow$ Generatemask

$$13 : \tilde{} \sim N(0, I);; \qquad t \sim U(0, 800)$$

$$14 : zt \leftarrow \sqrt{\alpha\acute{r}\, t}\, z0 + \sqrt{1 - \alpha\acute{r}t}\ \vartheta$$

15:$\epsilon \hat{} \leftarrow \epsilon\theta$ ([z_t; m; z_{masked}], t, e_f)

$$16 : L \leftarrow \| \epsilon\hat{} - \epsilon \|_2^2 / G$$

17: backward(L) with AMP scaling

18: if step mod G = 0 then

19: Clip gradients::$\|\nabla\theta\|$= 1.0

20: optimizer.step(); scheduler.step()

21: Save checkpoint if Lepoch<Lbest

22: end for

3.6 Inference with Classifier-Free Guidance

During inference, we employed classifier-free guidance [27] to control the tradeoff between sample quality and diversity. The guided noise prediction is:

Algorithm 2 describes the inference procedure.

Require: Trained pipeline θ, input sketch xs, prompt p, negative prompt pneg, guidance scale s,

steps SEnsure: Generated face image x^

1: es ← Es(xs); et ← CLIP(p)

2: e f ← F(es, et); e∅ ← CLIP(pneg)

3: m ← Interpolate(M(xs), h/8) ?Downsample mask to latent size

4: zT~ N (**0**,$\sigma_T{}^2$**I**) ? Initialize from noise

5: for t = T, T-1, . . . , 1 do

6: $\epsilon^{\wedge}_{uncond} \leftarrow \epsilon\theta$ ([z_t; m; z_{masked}], t, $e_{\emptyset}$)

7: ϵ^cond ← $\epsilon\ \theta$ ([z_t; m; z_{masked}], t, e $_f$)

8: $\tilde{\epsilon} \leftarrow \epsilon^{\wedge}_{uncond}$ + s · ($\epsilon^{\wedge}_{cond}$ - $\epsilon^{\wedge}_{uncond}$)

9: $z_{t\,1}$ ←1 Scheduler.step($\tilde{\epsilon}$, t, z_t)

10: end for

11: x^←clamp((z0/s_{vae})/2 + 0.5, 0, 1)

12: return x^

4. EXPERIMENTS

4.1 Dataset

We used the Person Face Sketches dataset [1] from Kaggle, which contains over 21,000 aligned sketch-photo pairs, to evaluate the model. TABLE 2 summarizes the dataset statistics.

Table 2: Person Face Sketches dataset statistics

Split	Sketches	Photos	Percentage(%)
Train	20,655	20,655	92.5
Val	1,000	1,000	4.5
Test	679	679	3.0
Total	22,334	22,334	100.0

Every couple is made up of a grayscale drawing of a hand-drawn face and color photograph. The preprocessing step resized all images to 256 × 256 pixels. Photos were scaled to the range of [0, 1],

and sketches were scaled to the range of [-1, 1] to be compatible with the VAE encoder. Horizontal flipping and random brightness/contrast jittering of data were data augmentation methods used during training.

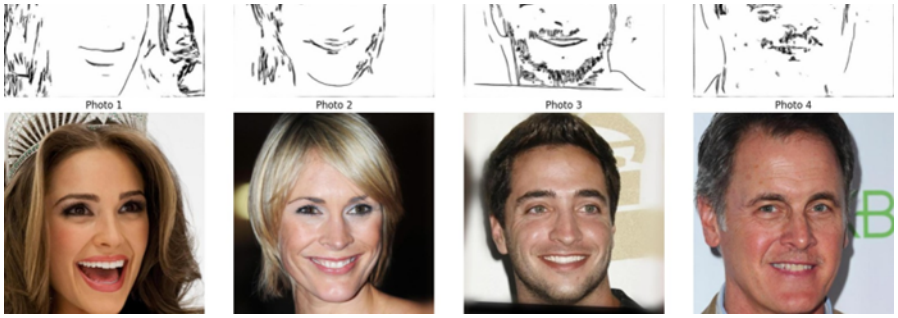

Figure 4: Examples of the Person Face Sketches.

Top row: input sketches. Bottom row: respective ground truth photographs (see FIGURE 4). The data is highly varied in terms of facial features, expressions, and style of sketches

4.2 Quantitative Results

The trained model was tested using a test set (679 pairs) with the following metrics (see TABLE 3):

1. Structural Similarity Index Measure (SSIM) : Compares structural similarity of generated and ground-truth images. SSIM had a range of $\in [0, 1]$; the higher the value, the better the structural preservation.
2. Learned Perceptual Image Patch Similarity (LPIPS): Measures the perceptual distance using deep features (AlexNet backbone). A lower LPIPS indicates higher perceptual similarity.
3. Fréchet Inception Distance (FID) : Measures the distance between feature distributions of real and generated images. Lower FID indicates better quality and diversity.

Table 3: Quantitative evaluation results on the test set

Metric	Value	Direction
FID	54.31	↓ (lower is better)
LPIPS	0.3785	↓ (lower is better)
SSIM	0.3041	↑ (higher is better)
Best Training Loss (MSE)	0.1630	↓ (lower is better)
GPU Memory (Training)	3.92 GB	↓ (lower is better)
Trainable Parameters	20.5 M	↓ (lower is better)

The SSIM value of 0.3041 reflects the complexity of the sketch-to-photo translation task, where the model must encode color, texture, and lighting details that are not provided in the input sketch. It is

essential to add that pixel-based measures of image quality like SSIM might not be indicative of the perceptual quality of the faces generated, and that between a given sketch and a valid photorealistic representation of that sketch there may be a line of valid representations [29]. All baselines were evaluated on the same test split, image resolution, and hardware configuration to ensure a fair comparison. For text-conditioned methods, the same prompt setting was used during inference as shown in TABLE 4.

Table 4: Benchmark comparison with GAN and diffusion baseline

Method	Model TYPE	SSIM↑	LPIPS↓	FID↓	ID-sim↑	Runtime↓	GPU-mem↓
pix2pixHD (sketch-only)	cGAN	0.3363	0.4896	45.68	0. 31	0.56	5.9
CycleGAN (sketch-only)	GAN		0.3378 0.4516	66.02	0.28	0.1393	13.5
ControlNet(sketch+text)	diffusion	0.1063 0.7301	75.78	0.44	2.078	6.91	
T2I-Adapter(sketch+text)	diffusion	0.0818	0.7436	82.5	0.47	1.46	6.32
Ours (text+sketch)	Cross-modal diffusion	0.3041	0.3785	54.31	0.5264	1.231	3.92

TABLE 4, reports a benchmark comparison between the proposed method and representative GAN-based and diffusion-based baselines evaluated under the same test protocol. The comparison includes structural similarity, perceptual similarity, distributional realism, identity similarity [30], inference time, and GPU memory usage. The results indicate that the proposed framework achieves a stronger overall trade-off among text-guided sketch-to-face diffusion baselines. Compared with ControlNet and T2I-Adapter, the proposed method improves structural similarity, perceptual similarity, distributional realism, identity consistency, inference time, and GPU memory usage. This supports the effectiveness of the proposed bidirectional sketch–text fusion and LoRA-based adaptation strategy. Compared with the sketch-only GAN baselines, the proposed method offers a better overall balance for text-guided sketch-to-face synthesis. It achieves the lowest LPIPS score of 0.3785, showing that the generated faces are perceptually closer to the ground-truth images. It also obtains the highest ID-sim score of 0.5264, which reflects stronger identity preservation. In addition, the method uses only 3.92 GB of GPU memory, making it more efficient than the compared GAN models. While pix2pixHD achieves a lower FID, it remains limited to sketch-only translation and does not include text guidance or identity-aware control. Therefore, the proposed framework is more suitable for the target task, as it combines perceptual quality, identity consistency, text conditioning, and memory efficiency in a unified framework.

4.3. Qualitative Results

FIGURE 5, presents representative sketch-to-face generation results for the test set.

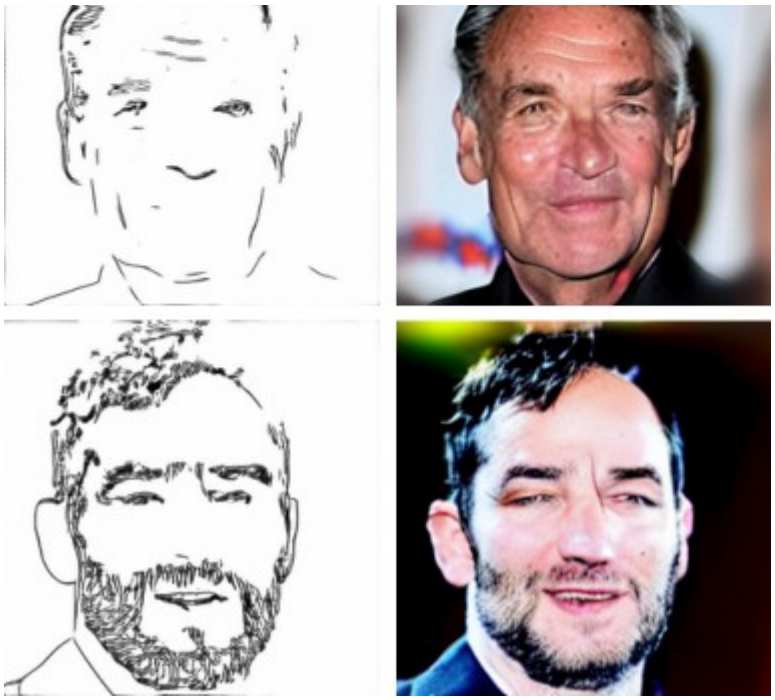

Figure 5: Test set qualitative results. Left column: input sketches.

Middle column: synthesized face images. Right column: ground truth photographs. The model preserves vital aspects of the faces including the shape of the face, the location of the eyes, shape of the nose and hairstyle and it gives a realistic skin texture and appropriate coloring.

Figure 6: The quality generation Publication of results showing the Sketch-to-Face pipeline can generate a wide range of faces of different degrees of detail and quality and of different styles of sketches.

We have several strengths in our approach which is reflected in its qualitative results:

1. Structural Fidelity: The produced outputs indicate a strong degree of geometric fidelity to the sketches, efficiently preserving facial structure, accurate eye localization, and the relative of nose landscapes.
2. Texture Synthesis The model has a good performance of producing realistic skin skin, realistic hair coloration, and realistic lighting conditions.
3. Diversity: With the aid of text prompts and structural consistency with the sketch, the Generation can be concerned with to the variety of ethnicity, lighting conditions, and styles (see FIGURE 5 and FIGURE 6).

4.4 Ablation Study: Classifier-Free Guidance Scale

We conducted an ablation study on the classifier-free guidance scale s to learn about its impact on the quality of image generation. FIGURE 7 shows the results for $s \in \{3.0, 5.0, 7.5, 10.0, 15.0\}$.

From left to right: input sketch and generated faces with $s \in \{3.0, 5.0, 7.5, 10.0, 15.0\}$. Low guidance ($s = 3.0$) produces blurry, low-fidelity results. The optimal range ($s = 7.5$) balances detail and naturalness. High guidance ($s = 15.0$) introduces over-saturation and artifacts[26].

Figure 7: Ablation study on classifier-free guidance scale.

4.5. Model Complexity Analysis

TABLE 5 presents a detailed breakdown of the model parameters and memory requirements.

Table 5: Parameter count and memory breakdown by component

Component	Parameters	Trainable	Status
VAE Encoder/Decoder	~83.7M	0	Frozen
U-Net (backbone)	~860M	0	Frozen
CLIP Text Encoder	~123M	0	Frozen
Sketch Encoder	1,778,528	1,778,528	Trainable
Cross-Modal Fusion	17,128,704	17,128,704	Trainable
Mask Generator	~10K	~10K	Trainable
LoRA Layers (128)	~18.7M	~18.7M	Trainable (FP32)
Total Trainable	20.5 M		
Frozen SD backbone	≈1.066B		
GPU Memory	3.92 GB Training		

The trainable parameters represent approximately 1.9% of the overall model parameters, indicating the effectiveness of the LoRA-based fine-tuning approach. This high memory footprint (3.92 GB) during training, combined with the fact that the approach can be trained directly on consumer-grade GPUs (e.g., NVIDIA T4 with 16 GB VRAM), implies that the barrier to entry is much lower than that of full fine-tuning approaches [31].

5. DISCUSSION

The diffusion-based framework has several benefits compared with its GAN-based counterparts. First, the iterative denoising process provides stability in training, preventing mode collapse and training instability, which are prevalent in GAN models. Second, the inpainting formulation can be extended to tasks such as the sketch-to-photo translation by considering the sketch as spatial guidance, enabling the diffusion process to infer missing details such as color, texture, and lighting. Third, guidance without classifiers provides a simple but efficient way to control the fidelity-diversity trade-off during inference without requiring model retraining.

Our proposed fusion module uses the bidirectional cross-attention design to achieve successful sketch-text integration. Unlike other models that concatenate or additively fuse text embeddings, the bidirectional attention can also have the text embeddings address spatial sketch features, and

vice versa, and the model can be trained to associate semantic concepts (e.g., realistic skin) with particular regions of geometry (e.g., facial contours). The 17.1M parameters in the fusion module indicate the most complicated parameter to train, which is the complexity of cross-modal alignment. Although the proposed method improves identity consistency compared with text-and-sketch diffusion baselines, further gains may be achieved by incorporating explicit identity-aware losses and stronger face-alignment preprocessing in future work.. Future directions include (1) scaling to higher resolutions, (2) including perceptual and identity losses, (3) scaling to multistyle sketch inputs.

6. CONCLUSION

This paper proposed Sketch to Face, an intermodal diffusion architecture for synthesizing photorealistic faces sketched by hand. We present a new sketch encoder that is capable of generating CLIP-wise embeddings, a cross-modal fusion unit with bilateral attention to combine sketches and text, and a memory-efficient training protocol using the LoRA fine-tuning algorithm. The results of experiments on the Person Face Sketches dataset demonstrate that the training convergence (minimal loss of 0.1630) is steady, structural similarity (SSIM = 0.3041) is plausible, and face appearance generation is qualitatively convincing. The quantitative evaluation further includes SSIM, LPIPS, FID, identity similarity, runtime, and GPU memory usage, providing a broader assessment of structural, perceptual, distributional, identity, and efficiency aspects. The model requires only 3.92 GB of GPU memory and 20.5 M trainable parameters, making it suitable for resource-constrained research settings. The ablation study on classifier-free guidance confirms that a scale of s = 7.5 provides the optimal balance between fidelity and naturalness. Our work demonstrates the viability of adapting pretrained diffusion models for specialized sketch-to-photo translation tasks through cross-modal conditioning and parameter-efficient fine-tuning

References

[1] Wang X, Tang X. Face Photo-Sketch Synthesis and Recognition. IEEE Trans Pattern Anal Mach Intell. 2009;31:1955-1967.

[2] Zhang W, Wang X, Tang X. Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition. In Conference on Computer Vision and Pattern Recognition. IEEE. 2011:513-520.

[3] Gao X, Wang N, Tao D, Li X. Face sketch–photo synthesis and retrieval using sparse representation. IEEE Transactions on circuits and systems for video technology. 2012;22:1213-1226.

[4] Saeed AJ, Hashim AA. Precise Gans Classification Based on Multi-Scale Comprehensive Performance Analysis. J Eng Sci Technol Rev. 2023;16:137-148.

[5] Isola P, Zhu JY, Zhou T, Efros AA. Image-To-Image Translation With Conditional Adversarial Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE. 2017:5967-5976.

[6] Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. InProceedings of the IEEE international conference on computer vision. IEEE. 2017:2223-2232.

[7] Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, et al. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE. 2018:8798-8807.

[8] Salim Rasheed A, Hasan Finjan R, Abdulsahib Hashim A, Murtdha Al-Saeedi M. 3D Face Creation via 2D Images Within Blender Virtual Environmen t. Indones J Electr Eng Comput. Sci. 2021;21:457-464.

[9] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, et al. Analyzing and Improving the Image Quality of StyleGAN. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE. 2020:8110-8119.

[10] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models. Adv Neural Inf Process Syst. 2020;33:6840-6851.

[11] Song J, Meng C, Ermon S. Denoising Diffusion Implicit Models. 2020. arXiv preprint arXiv: https://arxiv.org/pdf/2010.02502v1.

[12] Dhariwal P, Nichol A. Diffusion Models Beat Gans on Image Synthesis. Adv Neural Inf Process Syst. 2021;34:8780-8794.

[13] Saharia C, Chan W, Saxena S, Li L, Whang J, et al. Photorealistic Text-To-Image Diffusion Models With Deep Language Understanding. Adv Neural Inf Process Syst. 2022;35:36479-36494.

[14] Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical Text-Conditional Image Generation With Clip Latents. 2022. arXiv preprint: https://arxiv.org/pdf/2204.06125

[15] Lugmayr A, Danelljan M, Romero A, Yu F, Timofte R, et al. Repaint: Inpainting Using Denoising Diffusion Probabilistic Models. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE. 2022:11461-11471.

[16] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. Learning Transferable Visual Models From Natural Language Supervision. InInternational conference on machine learning. PmLR. 2021:8748-8763.

[17] Zhang L, Rao A, Agrawala M. Adding Conditional Control to Text-To-Image Diffusion Models. In Proceedings of the IEEE/CVF international conference on computer vision. 2023:3836-3847.

[18] Voynov A, Aberman K, Cohen-Or D. Sketch-Guided Text-To-Image Diffusion Models. InACM SIGGRAPH 2023 conference proceedings. 2023:1-11.

[19] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, et al. LoRA: Low-Rank Adaptation of Large Language Models. 2021. Arxiv preprint: https://arxiv.org/pdf/2106.09685.

[20] Micikevicius P, Narang S, Alben J, Diamos G, Elsen E, et al. Mixed Precision Training. In International Conference on Learning Representations. 2018.

[21] Chen T, Xu B, Zhang C, Guestrin C. Training Deep Nets With Sublinear Memory Cost. 2016. Arxiv preprint: https://arxiv.org/pdf/1604.06174

[22] Zhang M, Wang N, Li Y, Wang R, Gao X. Face sketch synthesis from coarse to fine. In Proceedings of the AAAI Conference on Artificial Intelligence 2018;32:1.

[23] Peng C, Gao X, Wang N, Li J. Face Recognition From Multiple Stylistic Sketches: Scenarios, Datasets, and Evaluation.Pattern Recognition. 2018;84:262-272.

[24] Canny J. A Computational Approach to Edge Detection. IEEE Trans Pattern Anal Mach Intell. 1986;8:679-698.

[25] Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, et al. Encoding in Style: A Stylegan Encoder for Image-To-Image Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:2287-2296.

[26] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis With Latent Diffusion Models. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022:10684-10695.

[27] Ho J, Salimans T. Classifier-Free Diffusion Guidance. 2022. Arxiv preprint: https://arxiv.org/pdf/2207.12598.

[28] Chen Y, Lai YK, Liu YJ. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:9465-9474.

[29] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:586-595.

[30] Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:815-823.

[31] Houlsby N, Giber A, Jastrzebski S, Morrone B, De Laroussilhe Q, et al. Parameter-Efficient Transfer Learning For NLP. In International Conference on Machine Learning. PMLR. 2019:2790-2799.